AN6100
R Analytics

FINAL- EXERCISES
R Analytics – Group A

Chin Chee Kai
cheekai@ntu.edu.sg
Nanyang Business School
Nanyang Technological University

Questions

In all questions, enter your results with all the decimal places which R provides into NTULearn.

Q1: A Covid-19 vaccination center in Singapore registers incoming people carefully with verification of identities and existing conditions. The registration times follow a normal distribution with a mean of 5 minutes and standard deviation of 4 minutes.

The number of people entering the center per minute follows Poisson distribution with an average arrival rate (λ) of 0.7 person per minute.

Using seed value 2398, set r1 to a vector of 60,000 random registration times. Set r2 to a vector of 60,000 random person-arrival-per-minute. Set r3 to be those registration times in r1 such that the corresponding value in r2 is positive.

How many values are in r3? Give the mean and sample standard deviation of r3.

Questions

In all questions, enter your results with all the decimal places which R provides into NTULearn.

Q2: A stock's daily closing price sequence can be described, in cents, by binomial distribution with n=100 and p=0.4 for the first 40 days. Subsequently, due to improvement in business and sentiments, the value of n increases by 15 every 20 days – thus for days 41 to 60 (inclusive), n=115, and for days 61 to 80 (inclusive), n will be 130.

Setting a seed value of 9877 (only once), create a vector prices which contains the first 200 days of stock's daily closing price sequence.

Give the mean and sample standard deviation of prices.

Suppose these 200 prices are future 200 days' actual stock prices. Knowing this, the stock trader wants to create a vector of "buy" and "SELL" instructions for the first 199 days. A "buy" is for today if the next day's price is strictly higher than today's price, and a "SELL" if next day's price is strictly lower than today's price; otherwise put a "-". Assume today's price is 50. Day 1 is tomorrow (ie the first future day). Create such a vector called instVec.

Give the number of "buy", and number of "SELL" in instVec.

What are the instructions on days 62, 112, and 190?

Questions

Q3: Using data file “AN6100-Data-3A.csv”, a reminder CSV file is to be produced to send reminders to those who have not gone for vaccination at all AND whose BMI is 22.0 and above. (NOTE: BMI is defined as weight/(height * height))

The CSV file just requires Lastname, Country, BMI and Age of those who need to be notified.

- How many entries are there in this reminder CSV?
- What is the Country of person whose Lastname is Jackson? (enter into NTULearn in exact same spelling and casing)
- How many entries are above 30 years of age?

Questions

Q4: Using data file “AN6100-Data-3A.csv”, perform divisive clustering only on the data set whose vaccination shots are 2, and only on columns Age, Height and Weight. Cut the dendrogram to make 3 clusters.

(a) Which cluster number is the smallest?

(b) How many people are in the largest cluster? How many in the smallest cluster?

(c) Assuming people in the CSV dataset are characterized by Age and Weight only, plot Age (X) vs Weight (Y) and optionally using clustering results to identify the points separately. Using only this resulting plot, assess the characteristics of people in the smallest cluster.

Write ONE LINE that best describes this smallest cluster.